DOCUMENT RESUME

ED 173 379 TM 008 991

AUTHOR Hughes, Francis P.
TITLE The Rasch Model Applied to the Equating of Several Examination Forms.
PUB DATE 11 Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Decision Making Skills; *Educational Assessment; *Equated Scores; *Graduate Medical Students; Inductive Methods; Item Sampling; Medical Research; *Multiple Choice Tests; Physical Examinations; Physician Patient Relationship; Test Construction; *Videotape Recordings
IDENTIFIERS *Rasch Model

ABSTRACT

In an examination designed by a medical specialty board, six tests were administered to groups of candidates for certification to investigate the feasibility of using videotapes of doctor-patient interactions to assess candidates' ability to note signs of potential physical disorders during a patient's physical examination and history taking, and their ability to draw reasonable conclusions and make appropriate decisions based on the available evidence. Common-item links, derived from the questions associated with selected interactions, were used to collect the data needed to equate the tests, and the Rasch model was the basis for scoring the tests and analyzing the equating data. Two indices were derived to help identify triads and links that lacked coherence. Investigation of the distributions of these indices is suggested so that firm guidelines for identifying triads and links that lack coherence can be established. Studies also should explore the relationship between coherent and incoherent triads and links, with the aim of clarifying the relationship and explaining the apparent contradiction detected in these data in which apparently coherent triads contained one or more seemingly incoherent links. (Author/1d)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED173379

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# THE RASCH MODEL APPLIED TO THE EQUATING OF SEVERAL EXAMINATION FORMS

FRANCIS P. HUGHES, National Board of Medical Examiners

APRIL 11, 1979

AERA/NCME JOINT SESSION

NUMBER 22.20

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Francis P. Hughes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM008 991

# THE RASCH MODEL APPLIED TO THE EQUATING OF SEVERAL EXAMINATION FORMS

FRANCIS P. HUGHES, National Board of Medical Examiners

## INTRODUCTION

Six tests were administered to different groups of examinees, and common-item links were used to collect the data needed to equate them. The Rasch model was the basis for scoring the tests and analyzing the equating data. The use of a method developed by Drs. Benjamin Wright and Ronald Mead to estimate the expected magnitude of the shifts between the tests is described in the paper, and its application is illustrated with an example in an appendix. Also, a procedure is suggested for evaluating the quality or coherence of the equating data, and its application to the data in this study is discussed.

## NATURE OF THE TESTS AND THE COMMON ITEM LINKS

The examination was designed by a medical specialty board to investigate the feasibility of using videotaped presentations to assess the observational skills and diagnostic acumen of its candidates for certification. With the consent of the participants, videotapes were made as physicians examined patients and questioned them about their medical history. A set of six multiple-choice questions was developed for each taped physician-patient interaction, and the tests contained either 60 or 66 questions involving 10 or 11 interactions between a physician and patient. The tests were administered to assess the candidates' ability to note signs of potential physical disorders during a patient's physical examination and history taking, and their ability to draw reasonable conclusions and make appropriate decisions based on the available evidence.

The examination format made it impractical to construct a common-item link between tests by sampling questions, so one or more interactions were chosen and the questions associated with them became the common-item link. Table 1 shows the composition of the six tests, A - F, and the size of the groups responding to them. Table 2 shows the composition of each common-item link, including the number of questions actually used to estimate the shift constant between the calibration scales of the two tests being joined by the common items.

Sampling the physician-patient interactions rather than the test questions may have influenced the coherence of the links. The coherence of the links also could have been affected by the possible lack of independence between the responses to questions based on the same taped interaction, or by the rather small size of some of the links. For these reasons the quality or coherence of the link network was evaluated.

## THE LINK NETWORK

Figure 1 illustrates the complete network of common-item links joining the six tests and shows the direction and magnitude of the observed link shifts[1], with their standard errors in parentheses. The observed link shifts can be presented in a matrix format as shown in Table 3, and this has proven to be a more useful way of viewing the data.

---

[1]The difficulty of a test's items can be shifted to the calibration scale of another test by adding the shift constant between the two tests to the calibrated difficulty for each question. If an indirect path is taken, the algebraic sum of the shifts required to travel the route is the constant to be added, with care taken to reverse the sign of the shift constant when the direction traveled is opposite that shown in Figure 1. Since the calibrated average difficulty of each test is 0.00, the shift constant also is the test's average difficulty on the new scale.

Table 3 also contains estimates of each test's average difficulty on a common scale and of the expected values of the shift constants between tests for which there were no common-item links. These estimates were obtained using an iterative procedure developed by Drs. Wright and Mead at the University of Chicago. First, the observed link shifts (including 0.00 for the shift of a test onto itself) were used to obtain row and column means. Then, the appropriate row and column means were added to estimate the shift for each empty cell in the matrix, i.e., the shift for each pair of tests that had no items in common. The process of averaging the row and column data and estimating the shifts for non-existent links was repeated until the row and column means stabilized. Appendix A illustrates the procedure with the data for this equating study.

The row means are estimates of the tests' average difficulty on a common calibration scale, and each sum of row and column means is an estimate of the expected value of the shift between two tests. In Table 3 the estimates of the expected values are shown only for the tests having no common items. The expected values for links having observed shifts are reported in Table 6.

Table 4 shows the average difficulty of each test when measured on its individual calibration scale (i.e., before equating) and on a common scale. Tests A, B, and C were roughly equal in difficulty with the estimates ranging from -0.27 logits to -0.13 logits. The difficulty estimates for tests D, E, and F ranged from 0.16 logits to 0.23 logits, showing them to be about equally difficult and generally harder than A, B, and C.

The mean ability estimates for the six examinee groups are shown

in Table 4 too, both before and after the tests had been equated. The estimates were placed on a common calibration scale by adding the difficulty of the equated tests to the ability of the groups as estimated on the individual calibration scales. For example, the average ability of group A was -0.27 + 1.51, or 1.24 logits. C was the most able group and the least able was clearly D. Group B was about .20 logits more able than D, and groups E and F were both about .10 logits more able than B. Group A was roughly midway between groups E and F and group C in ability.

## COHERENT LINKS AND TRIADS

The variable measured by a common-item link and each of the tests containing it should be the same if those tests are parallel and the link is a representative sample of each. If that is the case and the common items fit the model well, the link between the two tests should be coherent.

The link fit analysis described by Kreines and Mead (1979) excludes the poorest-fitting common items from the estimation of the link shift because their difficulty has changed significantly. To the extent the remaining common items do not fit well or the link and tests containing them measure somewhat different variables, the link lacks coherence or is incoherent.

A triad is any segment of the link network formed by three interconnected links, such as ACE, CDF, or ADE in Figure 1. The shift[2]

---

[2]The difficulty parameters for questions in one test can be shifted to the calibration scales of the other two tests forming a triad and returned to its own scale, by using the data from the equating analyses to circuit or traverse the triads.

needed to traverse a triad is the sum of the shifts for the links forming the triad, and its expected value is 0.00 logits since traversing a triad returns the item parameters to their original calibration scale. To the extent a triad shift deviates greatly from zero the triad lacks coherence. That is, one or more of its links lacks coherence, or two links relate differently to the same test and thus introduce incoherence into the triad.

Several authors (Doherty and Forster, 1976; Ingebo, 1976; Wright, 1977) have suggested the utility of the triad shift in identifying links that lack coherence. A triad with a large non-zero shift would represent a probable area of incoherence in the equating network, and the source of its incoherence would be identified by studying adjacent triads in the network. Since a link that lacks coherence should disturb the coherence of every triad containing it, at least one adjacent triad would be expected to lack coherence too. The link common to both incoherent triads would be the source of their incoherence.

This approach left unanswered the question of how large a triad shift is needed to signal a triad's lack of coherence, and in a complex link network the method suggested for identifying incoherent links is at best cumbersome if not impossible to implement. Therefore, a triad index and a link index were devised in an attempt to facilitate the evaluation of the quality of the equating data and the identification of triads and links that lacked coherence.

## THE TRIAD INDEX AND THE LINK INDEX

The observed and expected values of the triad shifts are known, and their standard errors ($SE_t$) can be estimated. Using these data, the triad indices were computed as follows:

$$I_t = (\text{Observed Shift} - \text{Expected Shift})/SE_t,$$

where $SE_t$ is equal to the square root of the sum of the variances of the three links forming the triad (i.e., $SE_t = \sqrt{\Sigma\ SE_l^2}$).

The observed values of the link shifts and their standard errors ($SE_l$) are known, and their expected values were estimated using the procedure described earlier. The link index was computed as follows:

$$I_l = (\text{Observed Shift} - \text{Expected Shift})/SE_l$$

The distributions of these indices have not been investigated, so they were used only to rank the triads and links and not to test hypotheses about their coherence. Unlike the triad index whose sign is arbitrary, the sign of a link index indicates whether the shift is over- or under-estimating its expected value, and it may be useful in the evaluation of a link's quality. The magnitude of the indices, however, indicated the relative coherence of the triads and links. Without knowledge of the indices' distributions no firm guidelines were available for determining coherence and incoherence, but the relative quality of the triads and links was ascertained and those that seemed most likely to be lacking coherence were examined.

Table 5 contains the triads ranked according to their triad indices. Triad CDF clearly has less coherence than the others, and the index for ADE also suggested it lacks coherence. It was not clear whether triads ACD and ACF lacked coherence, but the other four triads appeared to be relatively coherent.

Table 6 ranks the links according to the magnitude of their link indices. ED, DF, and FC seem to lack coherence and overestimate the magnitude of the shift. The DC link may or may not be coherent, but the other seven links have indices suggesting their coherence.

EXAMINING THE SOURCES OF INCOHERENCE

It was surprising to note that the least coherent links contained the largest number of common items and involved the most physician-patient interactions. The interactions were designed to assess similar skills and abilities in a variety of situations, so the content of the interactions was a sample of the more common situations the candidate could expect to encounter in independent practice. The specific nature of each interaction's content may be the cause of the incoherence noted in those links involving several physician-patient interactions.

The link and triad indices helped identify links ED, DF, and FC and triads CDF and ADE as lacking coherence. Triad CDF contained links DF and FC and this explained its lack of coherence. Similarly, link ED's incoherence provided a ready explanation for the lack of coherence noted in triad ADE.

DF, however, was part of triad ADF, one of the more coherent of the eight triads formed by the common-item links. This observation was not expected since it had been thought that an incoherent link would disturb the coherence of every triad containing it. Triad ADF also contained links DA and FA neither of which had unusually large link indices. The possibility that an incoherent link could be part of a coherent triad was pursued using other data from the equating analyses.

Link FC's lack of coherence did not disturb the coherence of triads BCF or ACF. The other links in both triads had indices suggesting they were coherent. This outcome supported the possibility that an incoherent link did not produce a lack of coherence in every triad containing it.

ED overestimated its shift since its link index was 2.25, but it

was part of CDE, a relatively coherent triad. The indices for DC (-1.50) and EC (-0.64) indicated less than serious underestimates of their link shifts, although the coherence of DC is uncertain. These data supported the contention that an incoherent link may not disturb the coherence of every triad containing it. They also introduced the possibility that links may offset one another's lack of coherence and give a triad the semblance of coherence.

## SUMMARY AND RECOMMENDATIONS

The equating of six tests was discussed using data obtained from common-item links and analyzed using the Rasch model. An iterative procedure developed by Drs. Wright and Mead for estimating average test difficulty on a common scale, and for estimating the expected value of shifts for common-item and non-existent links, was described and illustrated using these data. Tests A, B, and C were roughly equal in difficulty, but they were easier than tests D, E, and F which also were equally difficult. Group D was considerably less able than the other five groups, which differed in ability but by relatively small amounts. C was the most able examinee group.

Two indices were derived to help identify triads and links that lacked coherence, and they proved useful even though little is known about them. The distributions of these indices should be investigated so that firm guidelines for identifying triads and links that lack coherence can be established. Studies also should explore the relationship between coherent and incoherent triads and links, with the aim of clarifying that relationship. Such studies might help to explain the apparent contradiction detected in these data in which apparently coherent triads contained one or more seemingly incoherent links.

## REFERENCES


Doherty, V.W. and Forster, F. Can Rasch Scaled Scores Be Predicted from a Calibrated Item Pool? Paper presented to American Educational Research Association, San Francisco, 1976.

Ingebo, G. How to Link Tests to Form an Item Pool. Paper presented to American Educational Research Association, San Francisco, 1976.

Kreines, D.C. and Mead, R.J. Equating Tests with the Rasch Model. Paper presented to American Educational Research Association, San Francisco, 1979.

Wright, Benjamin D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-116.

# THE RASCH MODEL APPLIED TO THE EQUATING OF SEVERAL EXAMINATION FORMS

FRANCIS P. HUGHES, National Board of Medical Examiners

TABLE 1

SIZE OF TEST AND GROUPS

| | Test/Group | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C | D | E | F |
| Size of Each Test | | | | | | |
| Number of Physician-Patient Interactions* | 10 | 10 | 11 | 11 | 10 | 10 |
| Number of Items | 60 | 60 | 66 | 66 | 60 | 60 |
| Size of Each Group | 421 | 418 | 542 | 428 | 697 | 297 |

*The examinees viewed videotapes of physicians examining patients and taking their medical history. The questions were designed to assess the examinees' observational skills and diagnostic acumen.

TABLE 2

SIZE OF EACH COMMON ITEMS LINK

| | Size of the Link | | |
|---|---|---|---|
| | Number of | Number of Test Questions | |
| Link | Interactions | in the Link | Used to Estimate Shift |
| BC† | 1 | 6 | 3 |
| CA* | 1 | 6 | 5 |
| DA | 1 | 6 | 5 |
| DC | 1 | 6 | 6 |
| DF | 2 | 10 | 8 |
| EA* | 2 | 8 | 5 |
| EC* | 1 | 4 | 4 |
| ED | 3 | 9 | 9 |
| FA | 1 | 3 | 4 |
| FB† | 2 | 8 | 5 |
| FC† | 3 | 13 | 7 |

*4 questions (1 interaction) from the AC link were part of the AE link and were the entire CE link.

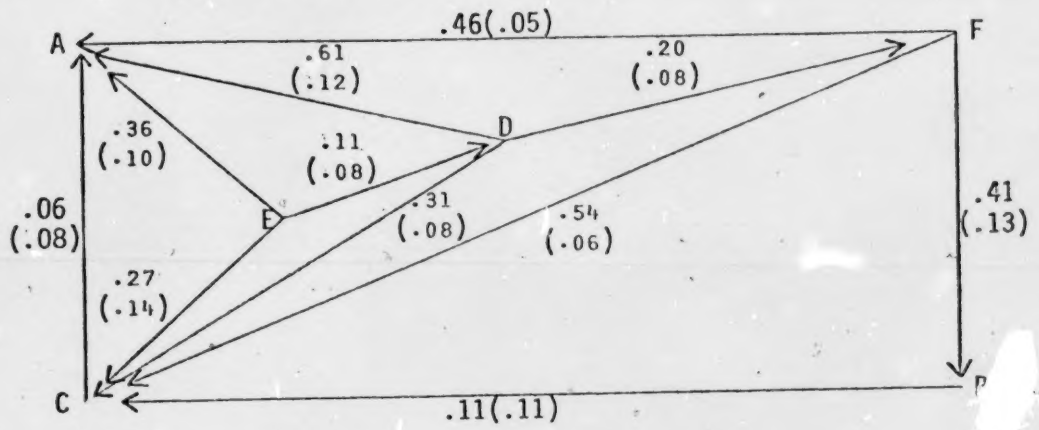
†5 questions (1 interaction) from the BC link were part of the BF and CF links.

THE RASCH MODEL APPLIED TO THE EQUATING OF SEVERAL EXAMINATION FORMS

FRANCIS P. HUGHES, National Board of Medical Examiners

FIGURE 1

THE COMPLETE LINK NETWORK*




*The observed shift for each link is shown, with the standard error in parentheses.

TABLE 3

OBSERVED AND EXPECTED VALUES OF EACH LINK SHIFT[1]

| Shift From Test | Shift to Test A | B | C | D | E | F | Average Test Difficulty[2] |
|---|---|---|---|---|---|---|---|
| A | .00 | [-.14] | -.06 | -.61 | -.36 | -.46 | -.27 |
| B | [.14] | .00 | .11 | [-.36] | [-.29] | -.41 | -.13 |
| C | .06 | -.11 | .00 | -.31 | -.27 | -.54 | -.20 |
| D | .61 | [.36] | .31 | .00 | -.11 | .20 | .23 |
| E | .36 | [.29] | .27 | .11 | .00 | [-.05] | .16 |
| F | .46 | .41 | .54 | -.20 | [.05] | .00 | .21 |
| Column Mean | .27 | .13 | .20 | -.23 | -.16 | -.21 | |

[1]The observed link shift is shown between tests containing common items. For tests with no common-items link, the expected value of the shift is shown in brackets. The expected value of the link shift for any cell in the matrix is the sum of the appropriate row and column means.

[2]The row means are estimates of the tests' average difficulty.

TABLE 4

TEST DIFFICULTY AND GROUP ABILITY

| Average Test Difficulty | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| before equating | .00 | .00 | .00 | .00 | .00 | .00 |
| after equating | -.27 | -.13 | -.20 | .23 | .16 | .21 |

| Average Group Ability | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| before equating | 1.51 | 1.19 | 1.55 | 0.64 | 0.99 | 0.97 |
| after equating | 1.24 | 1.06 | 1.35 | 0.87 | 1.15 | 1.18 |

TABLE 5

TRIAD SHIFTS*

| Triad | Triad Shift | Standard Error ($SE_T$) | Triad Index@ |
|---|---|---|---|
| BCF | 0.02 | 0.18 | 0.11 |
| ACE | 0.03 | 0.19 | 0.16 |
| ADF | 0.05 | 0.15 | 0.33 |
| CDE | 0.15 | 0.18 | 0.83 |
| ACF | 0.14 | 0.11 | 1.27 |
| ACD | 0.24 | 0.17 | 1.41 |
| ADE | 0.36 | 0.18 | 2.00 |
| CDF | 0.43 | 0.13 | 3.31 |

*The expected value of a triad shift is always 0.00.

@Triad Index $I_t$ = (Observed Shift - Expected Shift)/$SE_t$

TABLE 6

LINK SHIFTS

| Link# | Link Shift@ | Standard Error ($SE_l$) | Expected Value | Link Index@ |
|---|---|---|---|---|
| CA* | .06 | .08 | .07 | -0.13 |
| BC† | .11 | .11 | .07 | 0.36 |
| FA | .46 | .05 | .48 | -0.40 |
| FB† | .41 | .13 | .34 | 0.54 |
| EC* | .27 | .14 | .36 | -0.64 |
| EA* | .36 | .10 | .43 | -0.70 |
| DA | .61 | .12 | .50 | 0.92 |
| DC | .31 | .08 | .43 | -1.50 |
| FC† | .54 | .06 | .41 | 2.17 |
| DF | .20 | .08 | .02 | 2.25 |
| ED | .11 | .08 | -.07 | 2.25 |

#Each link name places the the more difficult test first. The direction of the link shift is from the harder to the easier test, so the shift constant always is positive.

@Link Index $I_l$ = (Observed Shift - Expected Shift)/$SE_l$. Positive indices indicate overestimates of the expected shift and negative indices underestimates.

*Four questions (1 interaction) from the AC link were part of the AE link and were the CE link.

†Five questions (1 interaction) from the BC link were part of the BF and the CF links.

## APPENDIX

### ITERATIVE PROCEDURE FOR ESTIMATING AVERAGE TEST DIFFICULTY AND THE EXPECTED VALUE OF LINK SHIFTS

| Iteration | Shift From Test | Shift to Test A | B | C | D | E | F | Row Means (Average Test Difficulty) |
|---|---|---|---|---|---|---|---|---|
| 0 | A | 0 | * | -.06 | -.61 | -.36 | -.46 | -.30 |
| | B | * | 0 | .11 | * | * | -.41 | -.10 |
| | C | .06 | -.11 | 0 | -.31 | -.27 | -.54 | -.20 |
| | D | .61 | * | .31 | 0 | -.11 | .20 | .20 |
| | E | .36 | * | .27 | .11 | 0 | * | .19 |
| | F | .46 | .41 | .54 | -.20 | * | 0 | .24 |
| | Column Means | .30 | .10 | .20 | -.20 | -.19 | -.24 | |

*These are links for which there were no common items. Estimates of their expected values are the sums of the appropriate row and column means and are shown in brackets in subsequent matrices.

| | | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|---|
| 1 | A | 0 | [-.20] | -.06 | -.61 | -.36 | -.46 | -.28 |
| | B | [.20] | 0 | .11 | [-.30] | [-.29] | -.41 | -.12 |
| | C | .06 | -.11 | 0 | -.31 | -.27 | -.54 | -.20 |
| | D | .61 | [.30] | .31 | 0 | -.11 | .20 | .22 |
| | E | .36 | [.29] | .27 | .11 | 0 | [-.05] | .16 |
| | F | .46 | .41 | .54 | -.20 | [.05] | 0 | .21 |
| | | .28 | .12 | .20 | -.22 | -.16 | -.21 | |

| | | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|---|
| 2 | A | 0 | [-.16] | -.06 | -.61 | -.36 | -.46 | -.28 |
| | B | [.16] | 0 | .11 | [-.34] | [-.28] | -.41 | -.13 |
| | C | .06 | -.11 | 0 | -.31 | -.27 | -.54 | -.20 |
| | D | .61 | [.34] | .31 | 0 | -.11 | .20 | .23 |
| | E | .36 | [.28] | .27 | .11 | 0 | [-.05] | .16 |
| | F | .46 | .41 | .54 | -.20 | [.05] | 0 | .21 |
| | | .28 | .13 | .20 | -.23 | -.16 | -.21 | |

| | | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|---|
| 3 | A | 0 | [-.15] | -.06 | -.61 | -.36 | -.46 | -.28 |
| | B | [.15] | 0 | .11 | [-.36] | [-.29] | -.41 | -.13 |
| | C | .06 | -.11 | 0 | -.31 | -.27 | -.54 | -.20 |
| | D | .61 | [.36] | .31 | 0 | -.11 | .20 | .23 |
| | E | .36 | [.29] | .27 | .11 | 0 | [-.05] | .16 |
| | F | .46 | .41 | .54 | -.20 | [.05] | 0 | .21 |
| | | .28 | .13 | .20 | -.23 | -.16 | -.21 | |